PREDICTION OF HEART DISEASES RISK USING NOVEL MACHINE LEARNING TECHNIQUES

Synopsis submitted in partial fulfilment of the

Degree of Doctor of Philosophy in Computer Science

By

Anuradha.P

17PHCSP007

Supervisor

Dr. (Mrs). Vasantha Kalyani David

Professor

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Educationfor Women

Coimbatore-641043

AUGUST 2024

1. INTRODUCTION

Heart diseases or Cardio Vascular Diseases (CVDs) are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. Coronary heart disease (CHD) occurs when the heart muscle's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries. If a coronary artery becomes completely blocked, it can cause a heart attack.

The risk factors causing heart disease that can be modifiable are high blood pressure, high cholesterol levels, smoking, diabetes, overweight or obesity, lack of physical activity, unhealthy diet and stress. Non-modifiable risk factors include age, sex, family history and race. By analysing the risk factors that lead to heart disease, medical practitioners would advise on lifestyle changes and if necessary, prescribe medicines to control those risk factors and prevent heart disease.

Machine Learning Models can be used for the early prediction of heart diseases based on the risk factors and thereby prevent them. These models would support medical practitioners in clinical decision-making.

2. RESEARCH MOTIVATION

The factors motivating this study are to identify risk of heart disease at an early stage and decrease the number of people who might develop heart illness. Also, to provide a support system to medical practitioners in decision making. This provides inspiration for academicians and researchers to develop a clinical decision-making support system that aid in early prediction of the illness. Building such systems would expand healthcare access in low-income countries at a reasonable cost.

ML models when used for heart disease prediction have the potential to bridge the gap between research and clinical practice. By translating research findings into practical tools, healthcare professionals can utilize these models as decision support systems.

3. STATEMENT OF PROBLEM

An effective technique for extending a patient's life is early diagnosis and prognosis. The lack of resources and the scarcity of medical practitioners in developing nations, where a substantial fraction of the humans live, is a serious problem that creates a bottleneck in this respect and causes heart disease to be diagnosed only at an advanced stage. Moreover, there is a scarcity of medical experts in rural areas to analyze clinical tests and provide expert opinions on the risk of heart disease. Most people approach medical practitioners after the onset of heart disease which results in low chances of recovery from the disease.

A system that predicts the risk of heart disease and advises seeking a medical expert’s opinion would address this problem. In this regard, numerous studies have been conducted to use machine learning techniques to predict cardiac disease. However, there is potential to explore improvements in the space of feature engineering, combining strengths of multiple models, etc., to make a new approach more robust.

Hence, research on developing ensemble ML models focusing on improved performance is highly sought after. Such enhanced heart disease prediction systems with improved accuracy would be of great support to clinical decision-making. The problem statement is "to develop performance enhanced heart diseases risk prediction system to support clinical decision-making".

4. AIM OF THE RESEARCH

To devise novel Machine Learning techniques towards developing an enhanced heart disease prediction system that supports clinical decision-making

5. OBJECTIVES OF THE RESEARCH

- To devise novel feature selection techniques to select relevant features for identifying the appropriate risk parameters for the prediction of heart disease with less computational cost
- To develop an ensemble classifier with high prediction accuracy by selecting an appropriate combination of diverse base learners
- To improve the ensemble model by optimizing the selection of classifiers in the ensemble model

6. CONTRIBUTIONS OF THE RESEARCH

- In the first stage of the research work, Feature selection by Feature Importance Scores of Gradient Boosting Algorithms is proposed with a significant reduction in the search space of the subsets of features.
- A novel wrapper feature selection method, ModifiedBoostARoota, is developed, which uses CatBoost as the base model and a novel feature elimination approach to efficiently identify the risk parameters of heart disease.
- In the second stage of the research work, A novel Super-Learner Ensemble Model (SLEM) with an appropriate combination of diverse classifiers, is proposed, to accurately forecast cardiac illness
- An Optimized Super Learner Ensemble Model (OSLEM), using the Whale Optimization Algorithm and pairwise divergence measure to select an optimal set of base learners, is devised in the third stage of the research work, to improve the efficiency of the classification model.

7. LITERATURE SURVEY

Wazery, Y. M. et al., [1] had designed an improved version of the slime mould algorithm (ISMA). ISMA is a randomization-based optimization technique and has lower computational complexity and avoids premature convergence in high dimensional datasets. On using K-Nearest Neighbors (KNN) classifier the accuracy obtained was 89.3%. Khemphila.A and Boonjing.V, [2], used Information gain and backward elimination for feature selection and Artificial neural networks scoring an accuracy of 92% on Cleveland dataset. J.Wang et al., [3] devised Stacking

model comprising of GNB, GB, RF,ET, ADB, MLP and XGB and achieved an accuracy of 90.7% on statlog dataset.

Using Infogain, Correlation, and ReliefF, Zahangir et al. [4] selected features for the Statlog dataset; ReliefF was scored lowest, hence it was left out of the final subset. The last two features from the feature ranking by RF using Classifier Attribute were omitted from the final subset of the SA heart dataset and the accuracy was 71% by using RF. S.Y. Hera et al., [5] used Random Forest (RF) for feature selection and multi-tier ensemble consisting of SVM, LR, KNN, ADA, GNB, GBC for classification achieving 84% on Statlog dataset. The MTE model's performance heavily relies on the quality and diversity of the training data. If the dataset is biased or incomplete, the model's accuracy may suffer. B. Kolukisa et al., [6] applied the linear discriminant analysis and the SVM algorithm, which yielded an accuracy of 92.74%. Kilic U et al., [7] used ABC algorithm and SMO on Z-Alizadehsani dataset and achieved 89.44% accuracy. Black Hole Optimization Algorithm (BHO) for feature selection and XGBoost for classification was employed by Rajadevi.R [8] and 92.3% was the accuracy on evaluating the model. On Arrhythmia heart dataset, Z. Arabasadi et al., [9] focused on the concept that CAD occurs if one of the left circumflex (LCX) or left anterior descending (LAD) or right coronary (RCA) arteries is stenotic [3]. By using hybrid Neural Network-Genetic algorithms model, those authors achieved 93.85% accuracy. Doppala, B. P. et al.,[10] demonstrated an accuracy of 88.4% on experimenting Ensemble(NB, RF, SVM, XGBoost) classifier with voting on Cleveland heart dataset. Mohammad-Ashraf Ottom & Walaa Alshorman [11], in their work, employed for each feature the ranks across the FS techniques namely Information gain, reliefF, gain ratio, oneR and symmetric uncertainty were summed up to get the accumulated rank. Features with rank<1 were eliminated. K-Nearest Neighbors (KNN) gave an accuracy of 78.90% which was higher compared to no Feature Selection [11]. Ensemble classifiers or hybrid classifiers were not experimented.

From the review of literature, it is observed that there is scope for Research work to improve feature selection technique for enhanced model performance and build classifiers focusing on improving prediction performance and model generalizability.

8. METHODOLOGY

Phase 1: Proposed work 1: Feature Selection using Feature Importance Scores by Gradient Boosting Classifiers and a significant reduction in the search space of the feature subsets

In this study, the feature importance scores by gradient boosting classifiers, XGBoost (XGB) and CatBoost (CatB), were obtained on heart datasets. Applying the forward selection method yields subsets of features using the FI score as the threshold. XGB, CatB and Hard Majority Vote Ensemble comprising of (CatB, GNB, LR, XGB, RF) are used for classification. The optimal feature subset which yields the highest accuracy is determined. The possible range of subsets that needs to be searched within the entire search space of feature subsets is determined.

Proposed Work 2: Proposed a novel wrapper feature selection method, ModifiedBoostARoota, which uses CatBoost as the base model and a novel feature elimination approach to efficiently identify the risk parameters of heart disease

In this study, a new wrapper feature selection approach called ModifiedBoostARoota (MBAR) is devised which is a modified version of the'BoostARoota (BAR)' algorithm (Chasedehan, 2018). MBAR differs from BAR by utilizing a different base model and using different feature elimination process in the algorithm. A comparison of MBAR and BAR algorithms is done by experimenting using both CatB and XGB as base models in each. Following the feature selection process, CatB and XGB are modeled on the selected set of features and experimental outcomes of MBAR are compared with contemporary methods of choosing features in terms of prediction performance on a number of heart disease datasets.

Phase 2: Proposed Work 3: Proposed a novel Super-Learner Ensemble Model (SLEM) with an appropriate combination of diverse classifiers, to accurately forecast heart disease

In this study, a number of ML classifiers,including the Support Vector Machine (SVM), LR, XGBoost (XGB), CatB, RF, KNN, DT, and Gaussian NB (GNB) are chosen. These classifiers are trained on the heart datasets via repeated stratified k-fold cross-validation. In contrast to more sophisticated statistical models, however, ensemble models cannot reliably draw valid conclusions about the predictors of interest. This problem is solved by the proposed Super Learner Ensemble Model (SLEM) which employs a meta learner to integrate predictions from multiple base models. The suggested ensemble model, SLEM, benefits from a new combination of base models that enhances its performance.

Phase 3: Proposed Work 4: An Optimized Super Learner Ensemble Model (OSLEM), using the Whale Optimization Algorithm and pairwise divergence measure to select an optimal set of base learners

The ideal combination of multiple models in an ensemble model can reduce bias, variance and overfitting, leading to better generalization and higher prediction accuracy. In this phase of the research, the Whale Optimization Algorithm (WOA) is utilized to select appropriate classifiers for the ensemble, hence enhancing the ensemble model. By taking into account both the accuracy of the combined classifiers and their pairwise diversity, the accuracy-diversity-based pruning strategy is introduced in thecontext of ensemble pruning.

In the proposed work, Initially, based on the diversity of the learning models, SVM, LR, GNB, RF, KNN, DT, MVE, XGB and CatB are chosen as base classifiers. WOA algorithm uses "whales" to indicate the existence or absence of key foundational classifiers. The WOA outputs a whale with good fitness value. This is repeated fordifferent sets of iterations. Then, the diversity measure value is calculated for each whale, and the most diverse whale is chosen as the base learner for the Optimized Super Learner Ensemble Model (OSLEM). The OSLEM ensemble classifier takes as input the combined predictions of all the base classifiers, and the meta learner, LR,

improves its performance by more effectively combining the results of eachof the base models. The overall methodology of this proposed work is shown in the following figure 1:

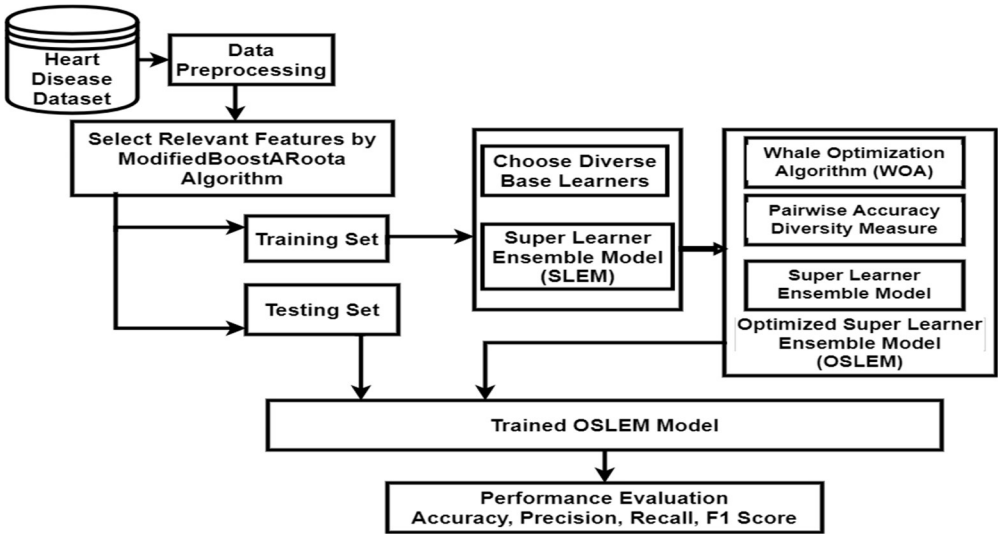

Figure 1: Overall Methodology

9. DATASETS USED

Seven datasets are used in this study are listed as follows:

i) Cardiovascular disease dataset taken from Mendeley data has 12 predictor attributes, 1 target and 1000 instances.

ii) Cleveland heart disease dataset taken from UCI Machine Learning Repository, has 303 instances and 13 predictor attributes.

iii) Statlog heart disease dataset taken from UCI Machine Learning Repository has 270 instances and 13 predictor attributes.

iv) South African heart disease dataset, hosted at Harvard Dataverse, has 9 features and 462 instances.

v) Cardiac Biomarkers dataset is real world data of patients who presented chest pain and had undergone lab tests, at Specialist hospital, Bangalore, India, in order to confirm whether they had a Myocardial Infarction (MI) or not. The personal details of the patients were not disclosed. The data collected consists of Age, Gender, CKMB, Myoglobin, Troponin-I, BNP, D-Dimer, ACS_types, Disease, shown in Table 1, consists of 192 instances.

v) Arrhythmia heart dataset in the UCI Machine Learning Repository, consists of ECG signals data with 279 attributes and 452 instances.

vi) Z-Alizadeh Sani dataset in the UCI Machine Learning Repository consists of 55 features related to coronary artery disease and 303 instances.

10. RESULTS AND DISCUSSION

In Proposed Work 1: Feature Importance (FI) Scores by using Gradient Boosting Algorithms is done. Each FI score is taken as threshold value, and feature subsets having FI > threshold is selected and accuracy of subsets of features are obtained by each classifier as shown in Table 1 and Table 2. We can see in all the above classification results in the four datasets, the highest accuracy of each of the classifier modeled on the datasets is around the median FI value. So, the

threshold range for searching for the optimal feature subsets can be 4 values above and below the median FI value which is 30% reduction in the search space of feature subsets.

Table 1: Performance Accuracy of classifiers on feature subsets

Feature Subsets	Cleveland HDS				Statlog HDS				South African HDS			
	MVE-Gain	MVE-PVC	CB-PVC	XGB-Gain	MVE-Gain	MVE-PVC	CB-PVC	XGB-Gain	MVE-Gain	MVE-PVC	CB-PVC	XGB-Gain
Top 1	78.69	75.41	75.10	78.69	77.78	72.22	72.22	74.075	66.67	68.82	77.74	67.74
Top 2	78.69	83.61	81.97	78.69	72.22	66.67	68.52	70.37	65.59	69.89	72.04	64.52
Top 3	86.89	85.25	86.89	86.69	77.78	**87.04**	**85.19**	72.22	70.97	74.19	72.04	72.04
Top 4	**88.52**	**90.16**	90.16	**88.52**	85.19	81.48	81.48	**85.19**	**73.12**	74.19	74.19	69.89
Top 5	85.25	88.52	**91.80**	83.61	79.63	83.33	79.63	83.33	69.89	**75.27**	77.42	**72.04**
Top 6	86.89	86.89	90.16	85.25	83.33	83.33	79.63	83.33	68.82	75.27	**78.49**	69.89
Top 7	86.89	86.89	88.52	85.25	85.19	81.48	79.63	83.33	69.89	75.27	76.34	72.04
Top 8	86.89	88.52	88.52	85.25	83.33	83.33	81.48	81.48	68.82	74.19	76.34	69.89
Top 9	86.89	88.52	88.52	83.61	83.33	85.19	83.33	83.33	69.89	75.27	75.27	68.82
Top 10	88.52	88.52	88.52	86.89	**87.04**	83.33	83.33	81.48				
Top 11	85.25	85.25	85.25	85.25	85.19	85.19	81.48	81.48				
Top 12	81.97	85.25	86.89	85.25	85.19	85.19	81.48	79.63				
All 13	81.97	86.89	86.89	85.25	85.19	83.33	83.33	77.78				

Table 2: Performance Accuracy of CatBoost , XGBoost and Majority Vote Ensemble classifiers on feature subsets of Z-Alizadehsani heart dataset

Feature Subsets withFI value >0.5 on Z-Alizadehsani heart dataset	XGB-Gain	MVE-Gain	MVE-PVC	CB-PVC
Top 1	85.06%	83.91%	83.91%	83.00%
Top 2	85.06%	83.91%	90.80%	87.00%
Top 3	91.95%	90.8%	88.51%	89.00%
Top 4	94.25%	93.1%	89.66%	89.00%
Top 5	94.25%	93.1%	91.95%	90.00%
Top 6	94.25%	93.1%	93.10%	90.00%
Top 7	94.25%	93.1%	93.10%	91.00%
Top 8	91.95%	91.95%	94.25%	92.00%
Top 9	91.95%	93.1%	93.10%	92.00%
Top 10	93.1%	94.25%	93.10%	92.00%
Top 11	94.25%	95.55%	93.10%	92.00%
Top 12	91.95%	95.4%	93.10%	91.00%
Top 13	91.96%	94.25%	93.10%	92.00%
Top 14	**95.4%**	95.4%	**95.40%**	92.00%
Top 15	94.25%	**96.15%**	94.25%	**93.00%**
Top 16 (Gain Median FI=2.77)	94.25%	**96.15%**	94.25%	92.00%
Top 17 (PVC Median FI=1.81)	**95.4%**	95.4%	**95.40%**	92.00%
Top 18	94.25%	95.4%	**95.40%**	92.00%
Top 19	95.4%	94.25%	93.10%	92.00%
Top 20	95.4%	95.4%	94.25%	92.00%
Top 21	94.25%	96.15%	93.10%	91.00%
Top 22	94.25%	96.15%	93.10%	92.00%
Top 23	94.25%	95.4%	94.25%	92.00%
Top 24	95.4%	96.15%	94.25%	92.00%
Top 25	95.4%	95.4%	95.40%	92.00%
Top 26	93.1%	95.4%	94.25%	92.00%
Top 27	94.25%	93.1%	93.10%	92.00%
Top 28	94.25%	94.25%	94.25%	92.00%

Top 29	94.25%	96.15%	91.95%	92.00%
Top 30	95.4%	93.1%	93.10%	92.00%

In Proposed Work 2, the proposed ModifiedBoostARoota (MBAR), a wrapper-feature selection algorithm, is a modified version of BoostARoota (BAR) algorithm (Chasedehan, 2018). In MBAR, CatBoost is used as base model whereas, BAR uses XGBoost as base model. MBAR has a new elimination procedure and the algorithm is as follows:

Algorithm ModifiedBoostARoota(MBAR):

1. Compute the shadow features (by randomly shuffling the original features) for each feature in the dataset and merge the shadow features with the original dataset to form an extended dataset of 'n' features.
2. Using any Tree-based model, compute the Feature Importance (FI) of all features in the extended dataset.
3. Assign rank to all the features, r_i , i=1 $to\ n$.
4. If the original feature's FI is less than the corresponding shadow feature's FI, then eliminate both the original and shadow feature.
5. If a feature's FI is insignificant then remove that feature.
6. Compute fscore of each feature in the extended dataset using, $fs_i = \frac{r_i}{FI_i},\ i = 1, \dots, n$
7. Compute weighted harmonic mean as: $whm = \frac{\sum r_i}{\sum fs_i},\ i = 1, \dots, n$
8. For any feature 'i' in the extended dataset, if $fs_i < whm$, then remove feature $'i\ '$ from the dataset.
9. If fs score of any feature is lesser than fs score of its shadow feature then eliminate it. Also, eliminate a feature if its fs score is negligible.
10. Repeat steps 1 to 9 until in each iteration atleast 10% of the features are eliminated, or if the maximum number of iterations has not been reached. Otherwise, return the remaining (chosen)features.

In Table 3, it is visible that MBAR with Catboost as base model and catboost as classifier gives the highest accuracy in all the datasets. Catboost classifier needs the least time to predict.

Table 3: Comparison of Performances of MBAR and BARon low dimensional datasets

Datasets	Feature Selection Algorithm & Classifier	Training time	Prediction time	Recall	Precision	Accuracy
Cleveland Heart Dataset	MBAR-CB, CBC	1.736 s	0.001 s	87%	96%	**91.8%**
	BAR-CB, CBC	1.954 s	0.002 s	87%	90%	88.52%
	MBAR-XGB, XGBC	0.034 s	0.004 s	87%	96%	91.8%
	BAR-XGB, XGBC	0.044 s	0.002 s	87%	84%	85.25%
SA Heart Dataset	MBAR-CB, CBC	1.331 s	0.001 s	48%	65%	74.19%
	BAR-CB, CBC	1.673 s	0.013 s	55%	59%	72.04%
	MBAR-XGB, XGBC	0.051 s	0.003 s	61%	63%	75.27%
	BAR-XGB, XGBC	0.084 s	0.002 s	58%	55%	69.89%
Statlog Heart Dataset	MBAR-CB, CBC	2.494 s	0.002 s	77%	95%	87.04%
	BAR-CB, CBC	2.626 s	0.003 s	81%	91%	87.04%
	MBAR-XGB, XGBC	0.032 s	0.003 s	73%	95%	85.19%
	BAR-XGB, XGBC	0.036 s	0.002 s	69%	95%	83.33%

The Table 4 shows the performance metrics of Catboost classifier after feature selection by MBAR are higher than that without feature selection.

Table 4 Performance Metrics of MBAR with Cat Boost onHigh-Dimensional Heart Datasets

Methods	Arrhythmia Dataset				Z-Alizadeh Sani Dataset			
	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
Without feature selection and with CatBoost classifier	84%	80%	82%	81.63 %	98%	92%	95%	95.38%
With MBAR selectedFeatures and CatBoost classifier	86%	81%	83%	**86.3%**	98%	94%	96%	**96.15%**

Thus, MBAR shows high performance on all the five datasets considered in the study.

InProposed Work 3: The proposed Super Learner Ensemble model (SLEM) comprising the base learners, Support Vector Machine (SVM), Logistic Regression (LR), XGBoost (XGB), CatB, RF, KNN, Decision Tree (DT), and Gaussian Naïve Bayes (GNB), are a combination of linear, probabilistic, bagging, boosting and stacking models. These classifiers are trained on the heart datasets by repeated stratified 10-fold cross-validation. The base learners are optimized using Backward elimination method and Catboost and DT are the optimal base learners of the SLEM model. Figure 2 depicts the SLEM classifier.

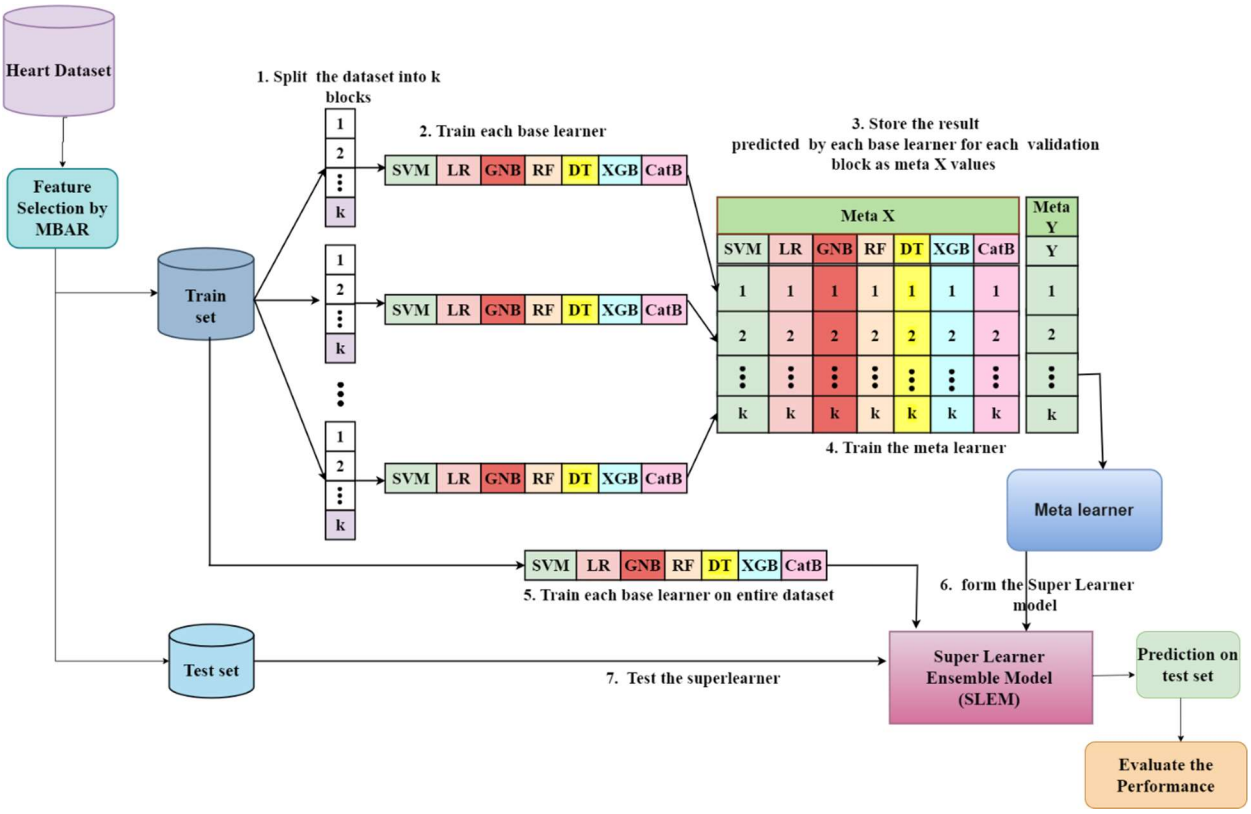
Heart Dataset
Feature Selection by MBAR
Train set
Test set
1. Split the dataset into k blocks
2. Train each base learner
3. Store the result predicted by each base learner for each validation block as meta X values
SVM LR GNB RF DT XGB CatB
Meta X
Meta Y
4. Train the meta learner
Meta learner
5. Train each base learner on entire dataset
6. form the Super Learner model
7. Test the superlearner
Super Learner Ensemble Model (SLEM)
Prediction on test set
Evaluate the Performance

Figure 2. Proposed Super Learner Ensemble model (SLEM)

In Proposed work 4, an Optimized SLEM (OSLEM) is devised where whale optimization algorithm and pairwise accuracy diversity measure is used to select an optimal set of base learners. Figure 3 depicts the design of the OSLEM which was described in the methodology section.

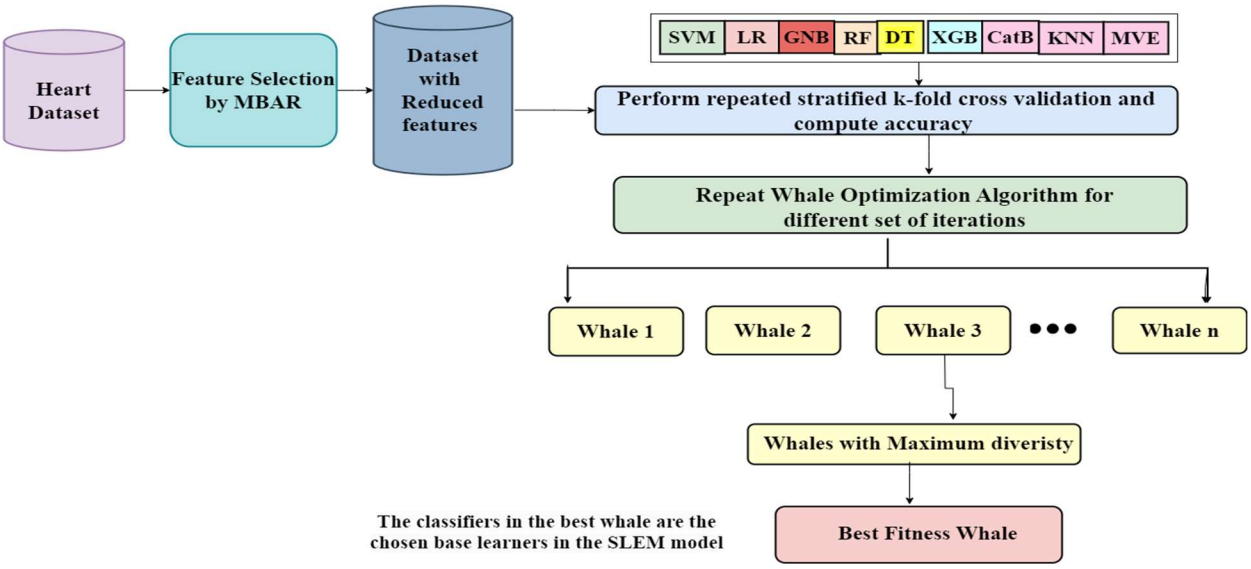

Figure 3: Optimized Super Learner Ensemble model (OSLEM)

The table 5 shows the comparison of OSLEM with SLEM and other classifiers on 5 different datasets with features selected by MBAR. The proposed OSLEM outperforms all other classifiers on all datasets. Figure 4 shows a visual representation of the comparison of accuracies obtained by all classifiers considered in this study on 5 low dimensional datasets. Figure 6 shows the graphical form of the performance accuracies of classifiers modelled on high dimensional datasets. We find OSLEM displaying the highest accuracy across all the seven heart datasets considered in this study.

Table 5. Performance comparison of OSLEM with other classifiers

Classifiers	**Accuracy on datasets (with Feature selection by MBAR) in%**				
	Cardiac Biomarkers Dataset	**Cleveland HD Dataset**	**SA HD Dataset**	**Statlog HD Dataset**	**Cardio-vascular Dataset**
CatBoost (CB)	100.0	85.5	81.8	83.3	99.0
Gaussian Naïve Bayes (GNB)	87.50	86.8	70.2	77.8	94.5
XGBoost (XGB)	100.0	84.2	76.9	84.4	97.2
Random Forest (RF)	100.0	85.5	76.8	83.3	98.6
Decision Tree (DT)	95.83	80.3	76.0	83.3	95.1
Majority Voting (MV)	93.75	88.2	71.9	83.3	98.3
Support Vector Machine(SVM)	75.0	85.5	72.7	77.8	79.7
K Nearest Neighbors (KNN)	81.25	82.9	72.7	83.3	83.1
Logistic Regression (LR)	93.75	86.8	69.4	77.8	96.9
Super learner Ensemble Model	97.9	85.5	81.0	83.3	98.6
SLEM with CatBoost & DecisionTree	**100**	**93.4**	**81.8**	**83.3**	**99.0**
Proposed Optimized Super learner Ensemble Model (OSLEM)	**100**	**93.4**	**89.2**	**92.6**	**99.0**

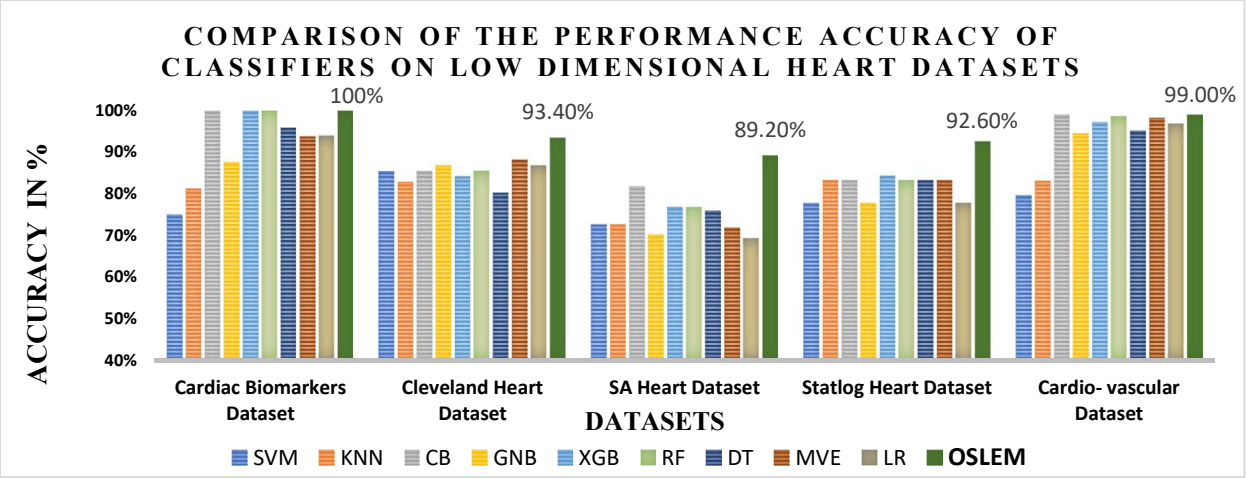

Figure 4: Comparison of performance Accuracy of classifiers

Table 6 displays the model evaluation metrics of the OSLEM on the low and high dimensional heart datasets. Results show that the OSLEM has high precision and has steadily achieved better metric values than other classifiers across all the seven heart datasets.

Table 6 Performance Metrics of OSLEM on Various Heart Disease Datasets

Dataset	Classifier	Precision	Recall	f1-score	Accuracy
Cardiac Biomarkers	MBAR+ OSLEM	100%	100%	100%	**100%**
Cleveland Heart	MBAR+ OSLEM	95.80%	88.50%	92.00%	**93.40%**
South African Heart	MBAR+ OSLEM	92.00%	74.10%	82.10%	**89.30%**
Statlog Heart	MBAR+ OSLEM	88.90%	88.90%	88.90%	**92.60%**
Cardio vascular disease	MBAR+ OSLEM	98.50%	99.20%	98.80%	**99. 0%**
Arrhythmia heart	MBAR+ OSLEM	88.00%	91.70%	89.80%	**90.00%**
Z-Alizadeh Sani heart	MBAR+ OSLEM	97.87%	97.87%	97.87%	**97.70%**

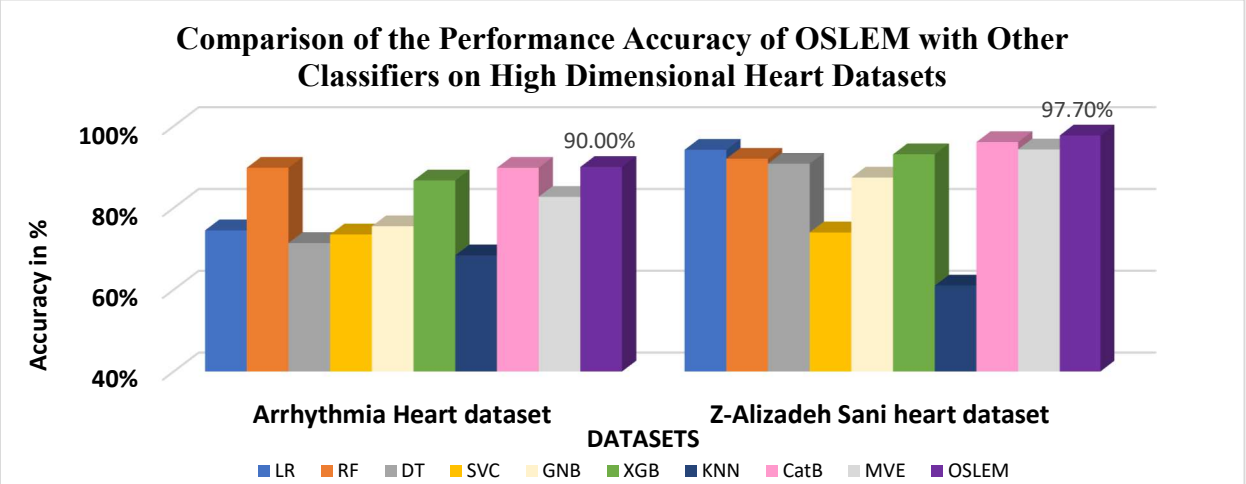

Figure 5: Performance Accuracy of classifiers on high dimensional datasets

Further, a comparative study of the proposed OSLEM model is done with other existing models in literature on cleveland, statlog, Arrhythmia and Z-alizadeh sani heart disease datasets. The proposed model has shown higher performance than the models considered in the study. Overall, the model evaluation metric values show that the proposed OSLEM model exhibits high performance on all the Heart Datasets considered.

11. CONCLUSION

- In Phase 1, Feature selection by Feature Importance (FI) score of Gradient Boosting algorithms proposes a significant reduction in the search space of the feature subsets compared to exhaustive search.
- A new wrapper feature selection algorithm, ModifiedBoostARoota (MBAR), with CatBoost as the base model and a novel feature elimination approach was devised in order to efficiently identify the risk parameters of heart disease. MBAR demonstrated high performance when evaluated on both low and high dimensional heart datasets.
- In Phase 2, a Super Learner Ensemble Model (SLEM) comprising SVM, LR, KNN, GNB, RF, DT, MVE, XGB, and CatB as base models was devised. The performance of SLEM improved with the base model combination of CatBoost and Decision Tree, identified by backward elimination.
- In Phase 3, an Optimized Super Learner Ensemble Model(OSLEM) is proposed in which the optimal combination of base classifiers in SLEM is determined using the Whale Optimization Algorithm (WOA) in conjunction with the pairwise disagreement diversity metric. OSLEM applied on heart datasets with features selected by MBAR outperforms the classifiers considered in this study.
- Proposed MBAR with OSLEM model outperforms other existing models on evaluating precision, recall, accuracy and f1- Score on both low and high dimensional heart datasets.
  A predictive system incorporating these ML techniques would serve towards early prediction of the risk of heart disease with an improved accuracy.

12. FUTURE WORK

- By combining information from various sources including health records, medical imaging, activity monitoring devices, and genomic data, researchers can gain deeper insights into the factors contributing to heart disease and develop more comprehensive prediction models.
- Using ML techniques, researchers can develop personalized risk assessment models which can optimize treatment plans and interventions, improving patient outcomes and reducing healthcare costs.
- Research experimenting cardiac biomarkers data of persons having family history of heart disease, might be a new pathway leading to early prediction of heart disease.

REFERENCES

[1] Wazery, Y. M., Saber, E., Houssein, E. H., Ali, A. A., & Amer, E. (2021). An efficient slime mould algorithm combined with k-nearest neighbor for medical classification tasks. IEEE Access, 9, 113666-113682.

[2] Khemphila.A and Boonjing.V. (2011). Heart Disease Classification Using Neural Network and Feature Selection. 21st International Conference on Systems Engineering, Las Vegas, NV, USA. pp. 406-409.

[3] J. Wang, C. Liu, L. Li, W. Li, L. Yao, H. Li, and H. Zhang. (2020). A stacking-based model for non-invasive detection of coronary heart disease. IEEE Access. Vol. 8. pp. 37124-37133.

[4] Mohammad Zahangir Alam, Saifur Rahman M., Sohel Rahman M. (2019). A Random Forest Based Predictor for Medical Data Classification Using Feature Ranking. Informatics in Medicine Unlocked, 15, 100180, 2019.

[5] S. Y. Hera, M. Amjad & M. K. Saba. (2022). Improving heart disease prediction using multi-tier ensemble model. Network Modeling Analysis in Health Informatics and Bioinformatics.11(41).

[6] B. Kolukisa et al. (2020). Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm. International Journal of Bioscience, Biochemistry, and Bioinformatics. 10(1).

[7] Kiliç Ü. and M. K. Keleş. (2018). Feature Selection with Artificial Bee Colony Algorithm on Z-Alizadeh Sani Dataset. 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-3.

[8] Rajadevi, R., Devi, E. R., Shanthakumari, R., Latha, R. S., Anitha, N., Devipriya, R. (2021). Feature selection for predicting heart disease using black hole optimization algorithm and xgboost classifier. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-7). IEEE.

[9] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard. (2017). Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. Comput Methods Programs Biomed. 141. pp. 19-26.

[10] Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., & Baik, N. (2022). A reliable machine telligence model for accurate identification of cardiovascular diseases using ensemble techniques. Journal of Healthcare Engineering, 2022, 1–13.

[11] Mohammad-Ashraf Ottom & Walaa Alshorman. (2019). Heart Diseases Prediction Using Accumulated Rank Features Selection Technique. Journal of Engineering and Applied Sciences, 14, 2249-2257.

[12] DeHan, C. (2018). BoostARoota. GitHub. https: // github.com /chasedehan / BoostARoota

[13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost:unbiased boosting with categorical features. Advances in neural information processing systems, 31, 1-11.

PUBLICATIONS

- Anuradha. P and Vasantha Kalyani David,"Improved Heart Diseases Risk Prediction using Optimized Super Learner Ensemble Model", International Journal of Intelligent Engineering and Systems, Vol.17, No.5, 2024, DOI: 10.22266/ijies2024.1031.25 **(scopus-indexed)**
- Anuradha. P and Vasantha Kalyani David, "Feature Selection by ModifiedBoostARoota and Classification by CatBoost Model on High Dimensional Heart Disease Datasets," *International Journal of Computer Theory and Engineering*, vol. 14, no. 4, pp. 141-148, 2022. doi: 10.7763/IJCTE.2022.V14.1321. **(scopus-indexed)**
- Anuradha. P and Vasantha Kalyani David, "Super Learner Model In Prediction Of Heart Attack Based On Cardiac Biomarkers", *Indian Journal of Computer Science and Engineering*, Vol. 12. No.6, 2021. doi: 10.21817 /indjcse/ 2021/ v12i6/ 211206076. **(Scopus-indexed)**
- Anuradha. P and Vasantha Kalyani David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using Gradient Boosting algorithms," *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 19-23, doi: 10.1109/ICCCIS51004.2021.9397154. **(paper is scopus indexed)(IEEE conference)**
- Anuradha. P and V. K. David, "Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS),* 2021, pp. 711-717, doi: 10.1109/ICAIS50930.2021.9395819. **(paper is scopus indexed)((IEEE conference)**
- Anuradha. P, Vasantha Kalyani David, "Feature Selection Using Whale Swarm Algorithm And A Comparison Of Classifiers For Prediction Of Cardiovascular Diseases", *International Journal Of Research And Analytical Reviews (IJRAR),* Volume.6, Issue 2, pp.123-130, June 2019, Available at HYPERLINK "https://www.ijrar.org/papers/IJRAR1ANP018.pdf" http://www.ijrar.org/IJRAR1ANP018.pdf
- Anuradha. P and Vasantha Kalyani David, "Leveraging The Power of Hybrid Machine Learning Algorithms to Predict Cardiovascular Diseases - A Review", i-manager's Journal on Computer Science,(2017), 5(3), 60-67. https://doi.org/10.26634/jcom.5.3.14018.